ICA graph

1 Experimental results

1.1 Deflation in Auddy and Yuan's algorithm.

In this section, we show that the deflation method in Auddy and Yuan's algorithm may not perform well. Auddy and Yuan's deflation step is as follows:

$$\mathbf{X}_j \leftarrow \mathbf{X}_i - ((\widehat{a}_j^{[t]})^\top \mathbf{X}_i)\widehat{a}_j^{[t]}, \ i = 1, ..., n \qquad \text{(deflation)},$$

where $\widehat{a}_j^{[t]}$ is estimated j-th column in mixing matrix through fixed-point iterations. After this deflation step, $\mathbf{X}_i$ is no longer a whitened data, which means we observe non-zero correlations and the variances of $\mathbf{X}_i$'s are away from 1). But Auddy and yuan's algorithm require the data to be whitened. If we whiten $\mathbf{X}_i$ again after the deflation, then $\langle \widehat{a}_j^{[t]}, \mathbf{X}_i \rangle \neq 0$. It seems that whitening and the orthogonalization are incompatible. Therefore given prewhitened data, we consider orthogonalizing unmixing matrix or mixing matrix in ICA model instead of orthogonalizing data $\mathbf{X}_i$.

Modifications from Auddy and Yuan's algorithm

1. We consider using sample fourth cumulant tensor(fourth moment tensor + M0) in Anadkumar(2014) or using Auddy's fourth moment tensor estimator combined with M0 in Anadkumar(2014) to estimate the fourth cumulant tensor.

2. We use the following orthogonalization :

$$\widehat{\mathbf{w}}_j^{[t]} \leftarrow \widehat{\mathbf{w}}_j^{[t]} - \sum_{m=1}^{j-1} < \widehat{\mathbf{w}}_j^{[t]}, \widehat{\mathbf{w}}_m^{[T]} > \widehat{\mathbf{w}}_m^{[T]}.$$

1.2 Data used in simulation

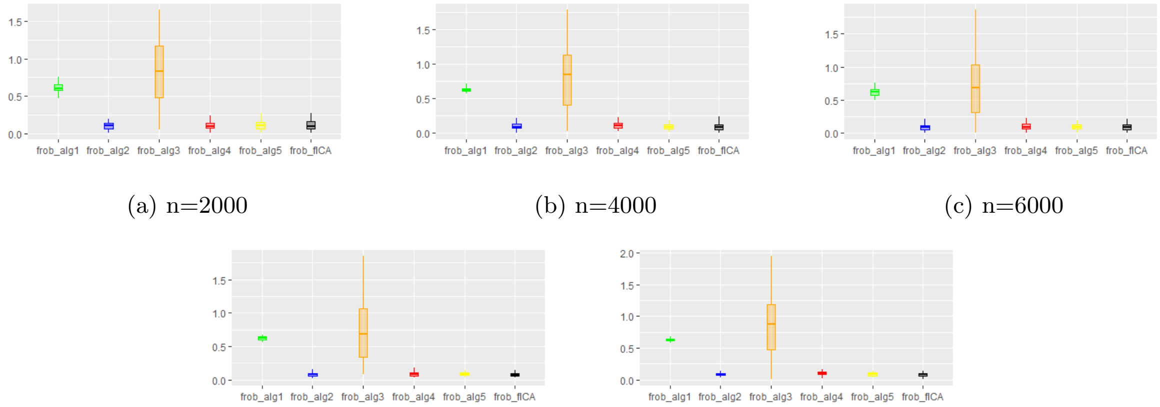
For the simulation study, we fixed sample size n=4000, and sources $\mathbf{Z}_i \in \mathbb{R}^d$ are independently generated from gamma distribution $\Gamma(j, 3), \forall 1 \leq j \leq d$, where j is the shape parameter and 3 is the rate parameter. Note that excessive kurtosis for gamma distributed random variable is 6/j. Unmixing matrix $\mathbf{W}$ was generated in a way that it contains one along its diagonal, -1/2 on superdiagonal, and 1/2 on subdiagonal. Then data was generated from the ICA model; $\mathbf{X}_i = \mathbf{W}^{-1}\mathbf{Z}_i, \ \forall 1 \leq i \leq n$.

1.3 Compare different M4hats and orthogonalizations

Figure 1 shows boxplots of relative errors, $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F / ||\mathbf{W}^*||_F$, from 50 times repeated experiment. We investigate five different settings and compare with R package "fICA" to test if **M4hat** and **deflation method** in Auddy and Yuan algorithm are valid. We denote **M4hat** as fourth cumulant tensor.

1. frob-alg1: relative errors when using **Auddy M4hat** and **Auddy deflation**

2. frob-alg2: relative errors when using **my M4hat** and **my deflation**
3. frob-alg3: relative errors when using **my M4hat** and **Auddy deflation**
4. frob-alg4: relative errors when using **Auddy M4hat** and **my deflation**
5. frob-alg5: relative errors when using **Auddy's fourth moment tensor combined with M0 from Anadkumar(2014)** and **my deflation**
6. frob-fICA: relative errors when using **R package "fICA"** with setting: kurtosis based fixed point iteration, and deflation scheme, FOBI or JADE estimator were used as initial estimate.



(a) n=2000

(b) n=4000

(c) n=6000

(d) n=8000

(e) n=10,000

Figure 1: Relative errors $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F / ||\mathbf{W}^*||_F$, d=5

	frob-alg1	frob-alg2	frob-alg3	frob-alg4	frob-alg5	frob-fICA
n=2000	30.60	5.54	41.35	5.49	5.49	5.39
n=4000	31.05	5.33	40.82	5.68	4.69	5.04
n=6000	31.12	4.48	35.53	5.20	4.72	4.62
n=8000	30.80	4.14	36.17	4.45	4.14	3.85
n=10,000	31.45	4.36	42.16	5.02	4.24	4.29

Table 1: Sum of relative errors over 50 iterations from Figure 1.

In figure 1 and table 1, it can be seen that frob-alg1 and frob-alg3 have large relative errors which implies that the deflation method in Auddy and Yuan's algorithm may not perform well. Also, when comparing frob-alg4 and frob-fICA(R package), Auddy and Yuan's initializer isn't performing better than FOBI or Jade initial estimate for kurtosis based fixed point iteration in deflation scheme.

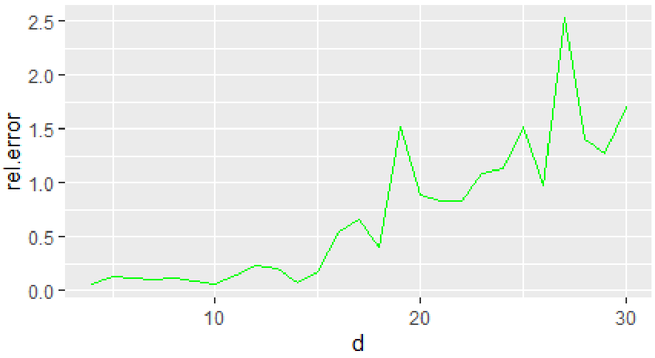
In next section, we write the modified algorithm, which corresponds to frob-alg2; use sample fourth cumulant tensor in Anadkumar(2014) and orthogonalize mixing matrix.

1.4 Modified algorithm

```
Algorithm 1: Kurtosis based fixed point iteration using random slicing for initialization
Data: $\mathbf{X}_1, ..., \mathbf{X}_n$, where $\mathbf{X}_j \in \mathbb{R}^d \ \ \forall 1 \le j \le n$.
1.Partition the samples into two halves with index sets $\mathbf{T}_1$ and $\mathbf{T}_2$ and center the variables.
2. Use eigenvalue decomposition to express the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \mathbf{PDP}^\top$ for
   $\{\mathbf{X}_i : i \in \mathbf{T}_1\}$, where $\mathbf{P}$ contains eigenvectors along its columns and $\mathbf{D}$ contains eigenvalues on its
   diagonal, and zeros elsewhere.
3. Prewhiten data from $\mathbf{T}_2 : \widetilde{\mathbf{X}_i} \leftarrow \mathbf{PD}^{-1/2}\mathbf{P}^\top \mathbf{X}_i$ for $i \in \mathbf{T}_2$.
4. Compute the fourth cumulant tensor estimate $\widehat{\mathcal{M}}$ by using prewhitened data from step 3:
Let $n_2 = n/2$ be the sample size of whitened data from step 3 and let $\widehat{\mathcal{M}_4} = \sum_{i=1}^{n_2} \mathbf{X}_i \circ \mathbf{X}_i \circ \mathbf{X}_i \circ \mathbf{X}_i$,
   where $\circ$ represents outer product. Also define a fourth-order tensor $\mathbf{K}$ as

   $$[\mathbf{K}]_{i_1,i_2,i_3,i_4} := \frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_1}x_{i,i_2}\right]\frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_3}x_{i,i_4}\right] + \frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_1}x_{i,i_3}\right]\frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_2}x_{i,i_4}\right] + \frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_1}x_{i,i_4}\right]\frac{1}{n_2}\left[\sum_{i=1}^{n_2} x_{i,i_2}x_{i,i_3}\right], \ 1 \le i_1, i_2, i_3, i_4 \le d.$$

Put $\widehat{\mathcal{M}} := \widehat{\mathcal{M}_4} - \mathbf{K}$.
5. for $j = 1$ to d do
       for $l = 1$ to $\mathbf{L}$ do                                  (Random slicing for initialization)
           a) Generate $d \times d$ matrix $\mathbf{G}$, where elements are randomly generated from
           standard normal $\mathcal{N}(0, 1)$.
           b) For each $1 \le i, j \le d$, compute $\mathrm{tr}(\widehat{\mathcal{M}}_{[\cdot,\cdot,i,j]}\mathbf{G})$ so that we have a $d \times d$ matrix.
           c) Compute the leading singular value and left singular vector of the matrix from step b),
           denoted by $\sigma_l$ and $\mathbf{u}_l$.
           d) Let $L^* := \mathrm{argmax}_{1 \le l \le L}\, \sigma_l$.
           Put $\widehat{w}_j^{[0]} := \mathbf{u}_{L^*}$.
       for $t = 1$ to $\mathbf{N}$ do                                  (Kurtosis based fixed-point iteration )
           a) $\widehat{w}_j^{[t]} \leftarrow \widehat{w}_j^{[t-1]} - \frac{1}{3n_2}\sum_{i=1}^{n_2}[\mathbf{X}_i((\widehat{w}_j^{[t-1]})^\top \mathbf{X}_i)^3]$
           b) If $j > 1$, do the following orthogonalization:

              $$\widehat{\mathbf{w}}_j^{[t]} \leftarrow \widehat{\mathbf{w}}_j^{[t]} - \sum_{m=1}^{j-1} \langle \widehat{\mathbf{w}}_j^{[t]}, \widehat{\mathbf{w}}_m^{[\mathbf{N}]} \rangle \widehat{\mathbf{w}}_m^{[\mathbf{N}]}.$$          (Deflation)

           c) Let $\widehat{w}_j^{[t]} \leftarrow \widehat{w}_j^{[t]} / ||\widehat{w}_j^{[t]}||_2$.
return: $\{\widehat{w}_j^{[\mathbf{N}]} : 1 \le j \le d\}$.
```

1.5 Relative errors in a function of d

We fix set.seed(1), n=4000 for this test. Figure 2 shows relative errors $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F/||\mathbf{W}^*||_F$ as a function of d.

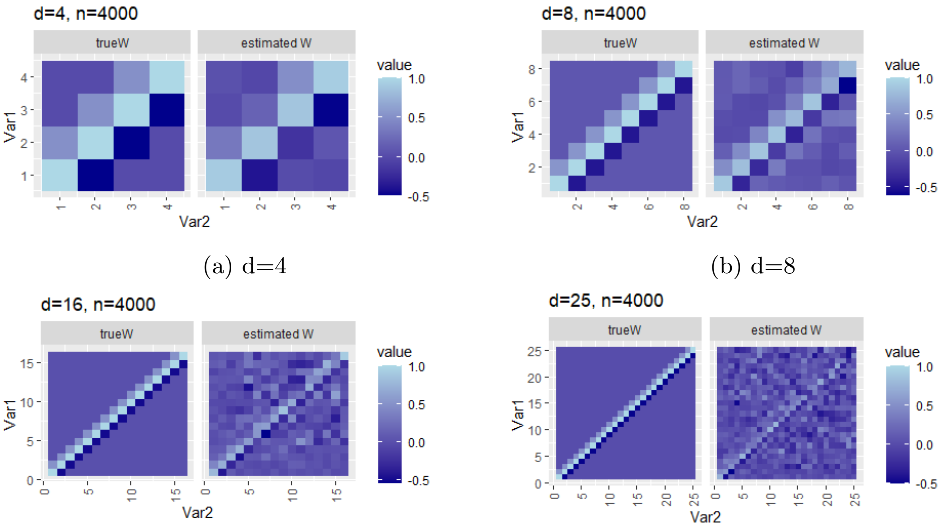


(a)

Figure 2: Relative errors $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F/||\mathbf{W}^*||_F$ for d=4,...,30

1.6 Heatmaps of Ws

Figure 3 shows heatmaps of true $\mathbf{W}^*$ and estimated $\widehat{\mathbf{W}}$ from the above algorithm 1 for d=4,8,16, and 25. We fixed the sample size n=4000.



(a) d=4

(b) d=8

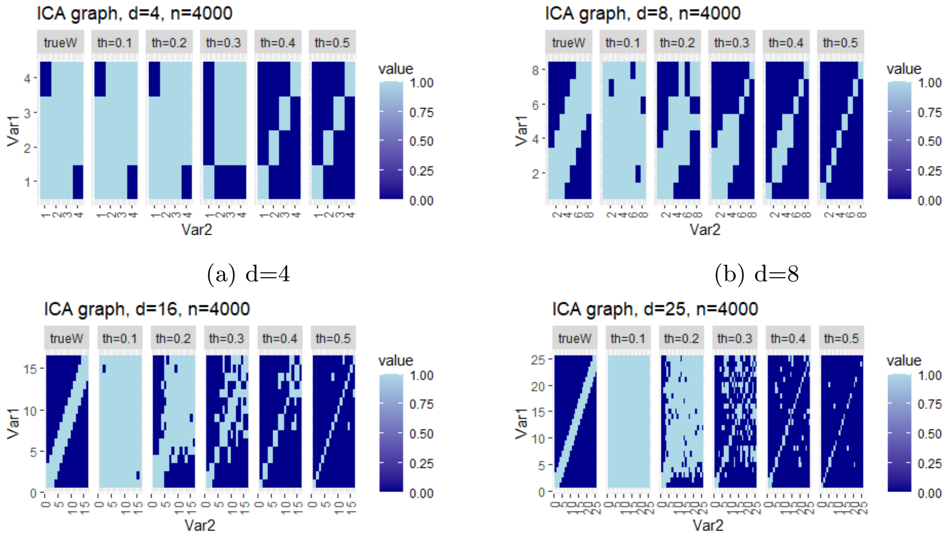
(c) d=16

(d) d=25

Figure 3: Heatmaps of true $\mathbf{W}^*$ and estimated $\widehat{\mathbf{W}}$, n=4000.

1.7 Heatmaps of induced graphs

By Proposition 2.2, we can construct a conditional independence graph from a given unmixing matrix $\mathbf{W}$. Figure 4 shows binary heatmaps representing conditional independence graph; i-th variable $\mathbf{x}_i^*$ and j-th variable $\mathbf{x}_j^*$ are conditionally independent if the i-th row, and j-th column area in a heatmap is dark blue and not conditionally independent if light blue. We fixed n=4000 and 5 different hard thresholds $HT = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ were applied on estimated unmixing matrix from algorithm 1; $\widehat{w}_{ij} = 1$ if $\widehat{w}_{ij} > t,\ t \in HT,\ 1 \leq i, j \leq d$ and zero otherwise.



(a) d=4

(b) d=8

(c) d=16

(d) d=25

Figure 4: Conditional independence graphs produced from $\mathbf{W}$'s, n=4000

1.8 Does initial estimate of W in algorithm 1 improves fastICA?

Again, from table 1, the initial estimate described in algorithm 1 (modified from Auddy and Yuan) is not better than Jade estimate or other initial estimate for the kurtosis based fixed point-iteration algorithm in deflation scheme.

Now we evaluate the initial estimator in algorithm 1. Figure 5 shows boxplots of relative errors, $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F/||\mathbf{W}^*||_F$, from 50 times repeated experiment, where $\widehat{\mathbf{W}}$ is the initial estimate of $\mathbf{W}^*$ produced from algorithm 1.

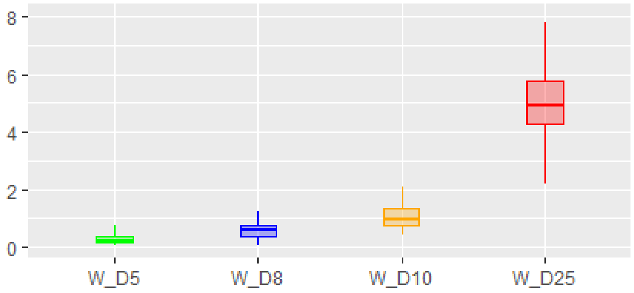
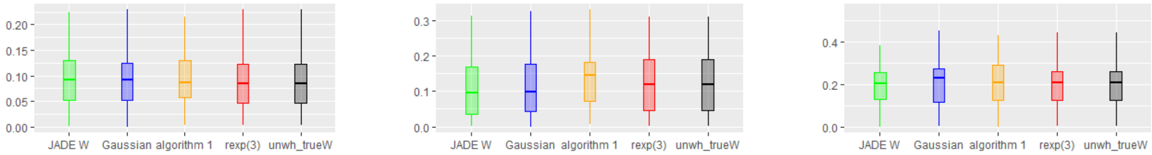


Figure 5: Relative errors $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F/||\mathbf{W}^*||_F$, n=4000

1.8.1 Investigate R package "fICA" with different initial estimates

Figure 6 shows boxplots of relative errors as above. We compare three different initial estimates for the kurtosis based fixed point-iteration algorithm in deflation scheme: 1) JADE estimator, 2) random matrix where entries are generated from standard normal, 3) Initial estimate produced from algorithm 1, 4) random matrix where entries are generated from expo(3), and 5) unwh-trueW= trueW $\% * \%\Sigma^{1/2}$, where σ is a covariance matrix of $\mathbf{X}$. Table 2 suggests the initial estimator produced from algorithm 1 may not be optimal for the kurtosis based fixed-point algorithm.



(a) d=5

(b) d=10

(c) d=15

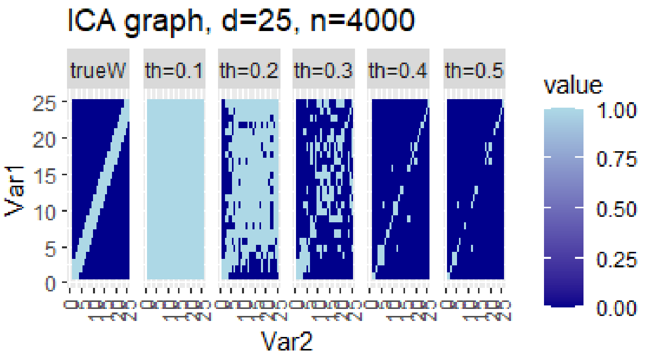
Figure 6: Relative errors $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F/||\mathbf{W}^*||_F$, n=4000

1.9 fICA on true $\mathbf{W}^*$

We can put the true W to initialize the fixed point algorithm using R package "fICA". Table 2 shows that this will not return a dramatically better estimate of true W. Figure 7 shows the sparsity of true W was put to "fICA" with diffrent thresholds: 0.1, 0.2, 0.3, 0.4, and 0.5. The heatmap seems similar to the one in figure 4.

	JADE W	Gaussian	alg 1	exp(3)	unwh-trueW
d=5	4.86	4.75	4.85	4.66	4.66
d=10	5.35	5.88	6.48	6.03	6.03
d=15	9.77	10.39	10.90	9.90	9.90

Table 2: Sum of relative errors over 50 iterations from Figure 6.



(a)

Figure 7: Conditional independence graph produced from true $\mathbf{W}^*$'s, n=4000

1.10 Can we improve 4th order cumulant tensor estimate?

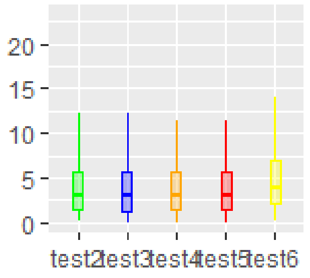
Let $m4 = \frac{1}{n}\sum_{i=1}^{n} x_i^4$, and $m2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, where x_i's are iid copies of single variate x_0 with $\mathbb{E}[X_0] = 0, \mathbb{E}[X_0^2] = 1$. We consider following schemes to estimate the fourth order cumulant tensor:

1. test2: $m4 - 3$
2. test3: $\frac{(n-1)^2}{n^2} \cdot m4 - 3$
3. test4: $\frac{(n-1)^2}{n^2} \cdot (m4/m2^2) - 3$
4. test5: $(m4/m2^2) - 3$
5. test6: $\frac{(n-1)}{(n-2)\cdot(n-3)} \cdot ((n+1)\cdot((m4/m2^2) - 3) + 6)$

Figure 8 shows boxplots of relative errors, $||\mathbf{W}^* - \widehat{\mathbf{W}}||_F / ||\mathbf{W}^*||_F$, from 200 times repeated experiment. We fixed n=4000, d=10 for the test. Test 2 has been used for algorithm 1 above, but table 3 implies that we may obtain a slightly more stable estimate of fourth order cumulant tensor by using "test4". For the details, see **Joanes, Gill(1998) "Comparing Measures of sample skewness and kurtosis".** But this simulation result depends on the distributions of $\mathbf{X}$, which are non-gaussian in this experiment.

	test2	test3	test4	test5	test6
d=10	3.8830	3.8771	3.8386	3.8387	5.0863

Table 3: Mean of relative errors over 200 iterations from Figure 8.



(a) d=10, n=4000, reps=200

Figure 8

https://cran.r-project.org/web/packages/PerformanceAnalytics/vignettes/EstimationComoments.pdf

1.11 Conclusions

The simulation result implies that

1. Initial estimate of mixing matrix produced from algorithm 1 is not competitive enough when compared to other initial estimate such as JADE estimator, see the table 2.

2. It seems that the kurtosis based fixed point algorithm for estimating unmixing matrix **W** is not very sensitive to its initializers.

3. Computation time for Algorithm 1 takes too long when d increases even for a single iteration.

what to do next

I would like to suggest we impose sparsity on mixing matrix A and approximate W by $W \approx A^\top(AA^\top)^{-1} = A^\top\Sigma$, where Σ is a covariance matrix of X. We may enforce sparsity on A by taking NMF(non negative matrix factorization + ICA assumption) model or assuming supergaussian sources such as Laplace distribution on ICA model. Then we may estimate sparse covariance which is well studied so that $W \approx A^\top\Sigma$ is sparse.

other thoughts on sparsity conditions

1. We cannot whiten X because we are interested in conditional independencies between variables in X. If X is not whiten, unmixing matrix is not orthogonal and mixing matrix $A = W^{-1}$, inverse of unmixing matrix W, is generally dense even if W is sparse.

2. We can consider sparse PCA, X=AS with L1 penalty on A, where A is loading matrix and S is principal component vector. Then A is orthogonal, so if A is sparse, then we have sparse $W = A^{-1}$. However, to apply The hammersley clifford theorem in our key lemma, we need S be multivariate gaussian so that uncorrelated principal components are independent. But we want to recover a conditional independence graph from non-gaussian components.

3. We may add non-negativity assumptions on both A and S in ICA model X=AS. Then we are only allowing additive combinations, so we may enforce A to be sparse. But still, without orthogonality of A, $W = A^{-1}$ is likely to be dense.

4. In ICA model X=AS, we want sparse $W = A^{-1}$. 1) Can we get full rank sparse A and enforce A^{-1} to be sparse as well? 2) What kind of dense A has sparse inverse?

5. Super-gaussian(for example, laplace): the probability density of the data is peaked at zero and has heavy tails (large values far from zero), when compared to a gaussian density of the same variance. Speech signals are usually highly supergaussian.

   If we assume super-gaussian sources, then we may enforce sparsity to sources, but negentropy(measure of non gaussianity) approach based on cumulant may perform very poorly; this is probably because it gives too much weight to the tails of the distribution, sensitive to outliers.less.Second, even if the cumulants were estimated perfectly, they mainly measure the tails of the distribution, and are largely unaffected by structure near the center of the distribution. This is because expectations of polynomials like the fourth power are much more strongly affected by data far away from zero than by data close to

   `https://www.cs.helsinki.fi/u/ahyvarin/papers/bookfinal_ICA.pdf`